AWARD NUMBER: W81XWH-14-1-0234

TITLE: Single-Cell RNA Sequencing of the Bronchial Epithelium in Smokers with Lung Cancer

PRINCIPAL INVESTIGATOR: Jennifer Beane-Ebel

CONTRACTING ORGANIZATION: Boston University School of Medicine
Boston, MA 02118-2340

REPORT DATE: July 2017

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE		Form Approved OMB No. 0704-0188

1. REPORT DATE	**2. REPORT TYPE**	**3. DATES COVERED**
July 2017	Annual	1JUL2016 - 30JUN2017

4. TITLE AND SUBTITLE
Single-Cell RNA Sequencing of the Bronchial Epithelium in Smokers with Lung Cancer

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-14-1-0234

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Jennifer Beane-Ebel

E-Mail: jbeane@bu.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Boston University School of Medicine
Department of Medicine
Division of Computational Biomedicine
Medical Campus
72 East Concord Street, E-631
Boston, MA 02118-2308

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Cigarette smoking, the major cause of lung cancer, creates a "field of injury" throughout the respiratory tract. We have previously shown that gene expression from bronchial epithelial cells reflects the physiologic response to cigarette smoke exposure and can serve as a diagnostic biomarker for lung cancer. The purpose of this Idea Development Award is to conduct single cell RNA sequencing on airway epithelial cells obtained from smokers with and without lung cancer to identify cell-type dependent gene expression alterations in the lung cancer field of injury.

Cells are being collected by brushing the right mainstem bronchus of smokers undergoing bronchoscopy for the suspicion of lung cancer. We have developed and optimized protocols to isolate single cells from these bronchial brushings using fluorescence-activated cell sorting (FACS) and to prepare libraries using an adapted version of the CEL-Seq RNA library preparation protocol that includes plate-, well-, and transcript-specific barcodes allowing hundreds of cells to be pooled together and sequenced. Additionally, we have developed computational pipelines to process the sequencing data into gene level counts for each cell as well as a new algorithm, Celda, to define and characterize transcriptionally distinct cell populations. We have successfully sequenced 3,456 cells collected by brushing the bronchial epithelium healthy never and current smokers and from high-risk current and former smokers with and without cancer. The data reveals the known and novel types of epithelial and immune cells. The results increase our understanding of our previously published gene expression changes associated with smoking, smoking cessation, and the presence of lung cancer in the bronchial airway epithelium.

Over the next year, we plan to sequence between 100 and 200 cells per donor from 6 healthy former smokers and 24 current and former smokers undergoing bronchoscopy for the suspicion of lung cancer. These samples will allow us to characterize the cell populations upon smoking cessation, in high-risk smokers without lung cancer, and in smokers with lung cancer. These data will provide a comprehensive single cell transcriptional profile of changes that occur in the bronchial epithelium in response to smoking, smoking cessation, and lung cancer. These discoveries may enhance current lung cancer diagnostics as well as suggest potential new therapeutics for lung cancer. Also, as this grant is in a no cost extension period, we have applied for NIH R01 funding to continue this project.

15. SUBJECT TERMS
single cell, bronchial epithelium, mRNA sequencing, and lung cancer

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			USAMRMC
Unclassified	Unclassified	Unclassified	UU	19	**19b. TELEPHONE NUMBER** *(include area code)*

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

Table of Contents

ABSTRACT:
Cigarette smoking, the major cause of lung cancer, creates a "field of injury" throughout the respiratory tract. We have previously shown that gene expression from bronchial epithelial cells reflects the physiologic response to cigarette smoke exposure and can serve as a diagnostic biomarker for lung cancer. The purpose of this Idea Development Award is to conduct single cell RNA sequencing on airway epithelial cells obtained from smokers with and without lung cancer to identify cell-type dependent gene expression alterations in the lung cancer field of injury.

Cells are being collected by brushing the right mainstem bronchus of smokers undergoing bronchoscopy for the suspicion of lung cancer. We have developed and optimized protocols to isolate single cells from these bronchial brushings using fluorescence-activated cell sorting (FACS) and to prepare libraries using an adapted version of the CEL-Seq RNA library preparation protocol that includes plate-, well-, and transcript-specific barcodes allowing hundreds of cells to be pooled together and sequenced. Additionally, we have developed computational pipelines to process the sequencing data into gene level counts for each cell as well as a new algorithm, Celda, to define and characterize transcriptionally distinct cell populations. We have successfully sequenced 3,456 cells collected by brushing the bronchial epithelium healthy never and current smokers and from high-risk current and former smokers with and without cancer. The data reveals the known and novel types of epithelial and immune cells. The results increase our understanding of our previously published gene expression changes associated with smoking, smoking cessation, and the presence of lung cancer in the bronchial airway epithelium.

Over the next year, we plan to sequence between 100 and 200 cells per donor from 6 healthy former smokers and 24 current and former smokers undergoing bronchoscopy for the suspicion of lung cancer. These samples will allow us to characterize the cell populations upon smoking cessation, in high-risk smokers without lung cancer, and in smokers with lung cancer. These data will provide a comprehensive single cell transcriptional profile of changes that occur in the bronchial epithelium in response to smoking, smoking cessation, and lung cancer. These discoveries may enhance current lung cancer diagnostics as well as suggest potential new therapeutics for lung cancer. Also, as this grant is in a no cost extension period, we have applied for NIH R01 funding to continue this project.

INTRODUCTION:
Cigarette smoking, the major cause of lung cancer, creates a "field of injury" throughout the respiratory tract by inducing molecular alterations such as allelic loss, p53 mutations, changes in promoter methylation and telomerase activity[1-5]. We have previously shown that gene expression from bronchial epithelial cells reflects the physiologic response to cigarette smoke exposure[6,7]. Importantly, we have extended this airway field of injury to the study of lung cancer, and have identified a bronchial airway gene expression signature that can serve as a diagnostic biomarker for lung cancer[8] that performs independently of clinical risk factors for disease[9]. The lung cancer diagnostic biomarker has been subsequently validated in in a large clinical trial[10] and has been commercialized by Veracyte, Inc. and is known as PERCEPTA™. Advances in technology for amplification of low amounts of RNA combined with next-generation sequencing have produced the ability to characterize the transcriptome of individual cells. While the bronchial brushings examined in our previous studies have captured a relatively pure population of bronchial epithelial cells, we are unable to discern which airway cell type or types are responsible for the gene expression changes observed nor characterize gene expression variation between cells. Variation in gene expression across single cells can be used to define unique subpopulations of cells that may be independent of known markers or cell morphology that may associate with lung cancer. We hypothesize that the lung cancer-specific gene expression in the bronchial epithelium might be restricted to specific known cell types (e.g. basal cells) or molecularly defined subpopulations of cells. The goal of this study will be to use single-cell RNA sequencing to identify cell-type dependent gene expression alterations in the lung cancer field of injury and to molecularly identify novel subpopulations of cells that are associated with lung cancer. These novel molecular insights hold the potential to improve the diagnostic utility of the airway epithelium for lung cancer and to guide new therapeutic strategies for lung cancer prevention.

KEYWORDS: single cell, bronchial epithelium, mRNA sequencing, and lung cancer

ACCOMPLISHMENTS:

What were the major goals of the project?

- Specific Aim 1: Identify which cell types in the airway epithelium harbor the lung cancer-specific alterations in biomarker genes by single cell RNA sequencing
    - Major Task 1: Isolate and sequence the RNA of single epithelial cells from the bronchus of smokers with and without lung cancer (n=15 subjects/group, n=960 single cells/subject).
        - Subtask 1: Approval of IRB and HRPO (1-2)

          The IRB at Boston University School of Medicine approved the study protocol on September 10, 2014 but notification of the outcome was received on December 4, 2014. The HRPO approval was obtained on December 19, 2014. This process took approximately 6 months to complete and delayed sample collection by about 4 months. Completion percent: 100%

        - Subtask 2: Collection of airway brushings from 30 subjects at BUMC (24-30)

          We have collected airway brushings from 28 never, current, and former smokers and 15 current and former smokers with and without lung cancer undergoing bronchoscopy for clinical suspicion of lung cancer. Of those subjects, we have adjudicated the cancer status of 12 of the 15 subjects with suspicion of lung cancer. We will continue to collect over the next year to increase the number of samples. Completion Percent: 85%

        - Subtask 3: Sorting of cells from brushings using FACS (Sorting of cells will take place within hours after collection) (24-30)

          Tissue acquired via bronchial brushings is dissociated, dead cells and red blood cells are excluded via FACS, and live cells are sorted into 96-well plates and frozen. This process needs to be completed immediately after sample collection. Therefore, we have sorted all samples collected above and completion is dependent on sample collection.
          Completion Percent: 85%

        - Subtask 4: mRNA isolation and library preparation (24-30)

          An established technique (CEL-Seq) for preparing single cell RNA-Seq libraries was adapted for this project and modified to increase sample multiplexing capacities and correct for experimental amplification biases. To date, we have produced high quality single cell data using this protocol and in the coming year we will process the samples collected. Completion Percent: 85%

        - Subtask 5: Sequencing of samples on Ilumina HiSeq 2500 (26-32)

          Single cell RNA libraries are massively multiplexed and paired-end sequencing is performed using the Illumina HiSeq 2500. Currently we have sequenced cells from 24 samples for a total of 3,456 cells. *Completion Percent: 85%*

        - Subtask 6: De-multiplex samples, preprocess, align, and analyze data quality (32)

          A computational pipeline developed in collaboration with the Yanai Lab (http://yanailab.technion.ac.il/) was used to preprocess and align reads generated via the CEL-Seq methodology. Additional metrics have been incorporated for the purposes of quality control and determination of sample cell type. The pipeline developed will be used to process the data as it is generated.

In addition, we spike-in ERCC controls into each well and compare the number of reads aligned to these controls to their known concentrations. A high correlation between read count and ERCC control concentration is on indicator of quality. Josh Campbell (co-investigator) has developed an algorithm known as celda that uses a Bayesian hierarchal model to discover novel transcriptional states. We have applied this algorithm in our analyses to gain interesting insights. Percent complete: 85%

- Milestone(s) Achieved: Generation of high quality single cell RNA sequencing data from 30 subjects

  After years of work, we have successfully created an experimental protocol that yields high-quality single cell sequencing data. Using this protocol, we have successfully generated high-quality data on 3,456 cells from 24 subjects. The data provides unprecedented insight into the cell populations present in the bronchial airway epithelium. We will use this protocol over the next year to complete sequencing the cells collected and sorted for additional subjects undergoing bronchoscopy for suspicion of lung cancer.

- Major Task 2: Determine the cell type(s) responsible for the aberrant gene expression in the airway lung cancer diagnostic biomarker. Completion Percent: 75%
  - Subtask 1: Summarize sequencing data into counts per gene (30)

    We have implemented a computational pipeline to summarize the sequencing data into counts per gene. We have gene counts for all 3,456 cells sequenced to date. Percent complete: 85%

  - Subtask 2: Classify gene expression as signal or noise based on a mixture model (30)

    An average of ~1500 genes were detected per cell and genes with less than 2 detected transcripts in 5 cells were excluded from further analyses. Percent complete: 85%

  - Subtask 3: Determine cell types of origin for lung cancer biomarker genes (30-34)

    We have defined transcriptionally distinct cell populations using Dr. Campbell's novel algorithm known as celda that leverages Latent Dirichlet Allocation (LDA) to determine the probability that sets of co-expressed genes (interpreted as "transcriptional states") are expressed by specific cellular subsets. Populations and their corresponding transcriptional states were visualized using the t-SNE (t-distributed Stochastic Neighbor Embedding) dimensionality reduction approach. We have conducted this analysis using the data from the 3,456 cells. We have subsequently examined expression of our lung cancer biomarker genes in these populations. Percent complete: 85%

  - Subtask 4: Identify cell type dependent lung cancer associated differential expression using a linear modeling and ANOVA strategy (32-34)

    We have identified smoking-associated cell type specific differential expression in the cells sequenced from never and current smokers. We will implement these same strategies to identify lung cancer-associated cell type-specific genes over the next year as we further analyze the sequenced samples. Percent complete: 28%

        - Subtask 5: Choose candidates for validation (34)

          We will be obtaining normal biopsies from the bronchial airway from current and former with and without cancer. We plan validate our findings using immunofluorescence to stain for genes expressed by lung cancer specific cell populations.

        - Milestone(s) Achieved:
          We have identified known lung cancer specific gene expression alterations that we observe to have cell type specific expression upon projection into our single cell data. Sequencing of samples from lung cancer subjects over the next year will confirm these observations and likely identify additional genes and cell populations.

- Specific Aim 2: Identify unique cell populations in the airway of smokers that are associated with lung cancer. Completion Percent: 45%
    - Major Task 1: Molecularly identify subpopulations of cells irrespective of cell type and determine if these cells are more or less abundant in the airways of patients with lung cancer
        - Subtask 1: Identification of novel subpopulations of cells using both class discovery and pathway prediction approaches (34-36)

          We have used celda to identify transcriptionally distinct cell populations and genes expressed by these populations and are characterizing the functions (pathways) of each population. We will implement these same strategies to identify lung cancer-specific populations of cells over the next year as we process the collected samples.

        - Subtask 2: Identify subpopulations associated with lung cancer

          We have identified smoking-associated subpopulations using celda. We will implement these same strategies to identify lung cancer-specific subpopulations over the next year as we process the collected samples.

        - Subtask 3: Choose candidates for validation (34-36)

          We are continuing to validate via immunofluorescence shift in cell type abundance and the existence of a novel cell type in the never and current smoker dataset. This work is a foundation for how we will validate the finding in the subjects with cancer.

        - Milestone(s) Achieved: Identification of novel subpopulations of airway epithelial cells that are associated with lung cancer (34-36)

          We have identified subpopulations of airway epithelial cells in 1,140 cells collected from never and current smokers. We have characterized the expression patterns of our previously defined lung cancer-specific biomarker genes in this dataset. This dataset is an important milestone in understanding the data that we are generating from current and former smokers with and without cancer. In order to evaluate changes associated with cancer we have to first understand the underlying differences that occur with smoking. In the next year, we plan to sequence cell from healthy former smokers to understand what changes revert upon smoking cessation. We can apply these findings to the interpretation of our cancer/no cancer data. We have recently obtained sequencing from 12 cancer/no cancer subjects and are currently analyzing the

data and have early evidence that novel subpopulation of cell may be present.

- Major Task 2: Validate lung cancer associated genes from specific cell types or from novel subpopulations in bronchial epithelial cells from independent subjects (n=10) using FISH
  - Subtask 1: Collect airway brushes from 10 subject for validation (24-36)
  - Subtask 2: RNA-FISH will be used to validate 5 candidate genes in conjunction with 4 known epithelial marker genes (30-36)
  - Milestone(s) Achieved: Validation of novel lung cancer-associated gene candidates in specific populations of epithelial cells

    We are validating our findings based on the cells sequenced from healthy never/current smokers via immunofluorescence on lung tissue obtained from never/current smokers. We are validating the observed shift in secretory cell type abundance and the existence of a novel cell type. We will use this same methodology to validate the findings among the cells sequenced from cancer/no cancer subjects instead of RNA-FISH. Additionally, we plan to validate the results using bronchial biopsies so that structure of the epithelium is intact and we can see spatial localization of our signals. This would not be possible using airway brushes.

What was accomplished under these goals?

The major objective is to conduct single cell RNA sequencing on cells collected from bronchial brushings from current and former smokers with and without lung cancer. In order to attain this goal, we developed and optimized the experimental protocols for FACS sorting, library preparation, and single cell data analysis. We have developed an unbiased methodology for sorting the cells from bronchial brushings into 96-well plates, a library preparation protocol that will produce high quality data, and an analysis pipeline for processing the data. As a result of the significant challenges in producing high-quality data, we elected to first study bronchial epithelial cells from never and current smokers. This follows our previous work of characterizing smoking-associated gene expression changes in the bronchial airway epithelium prior to our focus on lung cancer[6,7]. The smoking-associated gene expression changes found by microarray in our previous studies are very robust and reproducible and represented an ideal system for optimizing our protocol. We have generated high-quality data on 6 never and 6 current smokers by sequencing 1,140 cells (**Table 1**).

Smoking status	Age	Sex	Pack years
Never smoker (n=6)	30 (7.4)	3M, 3F	0 (0)
Current smoker (n=6)	43 (9.8)	3M, 3F	15.3 (7.9)

Table 1. Phenotypic information on healthy never and current smokers.

scRNA-Seq of Airway Cells from Disease-free Never and Current Smokers Reveals Known Epithelial Cell Types and Smoking-associated Cell Population Shifts: Bronchial brushings were obtained via volunteer bronchoscopy from 6 never smokers and 6 current smokers. For each donor, single $ALCAM^+$ epithelial cells or $CD45^+$ white blood cells (WBCs) were sorted into 96-well PCR plates containing a lysis buffer. scRNA-seq of human bronchial airway cells was performed using the CEL-Seq RNA library preparation protocol[11] and sequenced on the Illumina HiSeq 2500. ERCC spike-in RNA was added to every well to serve as a positive control and one well was left empty as a negative control. After aligning all reads to the genome for each cell (n=1,140) we quantified the expression of all protein-coding genes (between 1000-2000 genes were detected on average per cell within each donor.). We used a dimensionality reduction approach called t-Distributed Stochastic Neighbor Embedding (t-SNE) to visualize the relationship amongst cells (**Figure 1**). Never and current smoker cells were highlighted in grey and black, respectively (**Figure 1A**). Furthermore, relative expression of known airway epithelial marker genes was denoted in red (KRT5 = basal cells) (**Figure 1B**), yellow (FOXJ1 = ciliated cells) (**Figure 1C**), blue (MUC5B = secretory cells) (**Figure 1D**), and green (MUC5AC = goblet cells) (**Figure 2E**). Visual inspection of these plots indicates known subsets of airway epithelial cells are clearly present and transcriptomically distinct. The data indicates that MUC5AC+ goblet cells are more abundant in current smokers, whereas basal and MUC5B+ secretory cells are more abundant in never smokers.

Interestingly, when examining protein level expression of MUC5AC and MUC5B by immunofluorescence microscopy in large airway tissue obtained from an independent cohort of never (n=5) and current (n=5) smokers, it was found that in never smokers, there are clear subsets of MUC5B$^+$ and MUC5AC$^+$ cells, along with cells that are double positive for MUC5AC and MUC5B (**Figure 1F**). However, in the airways of current smokers, cells either only expressed MUCAC or were double positive for MUC5AC and MUC5B (**Figure 1F**). Given that it has been purported that MUCAC and MUC5B play distinct roles in airway homeostasis, this may be indicative of a major functional shift amongst secretory cells in response to cigarette smoke[12].

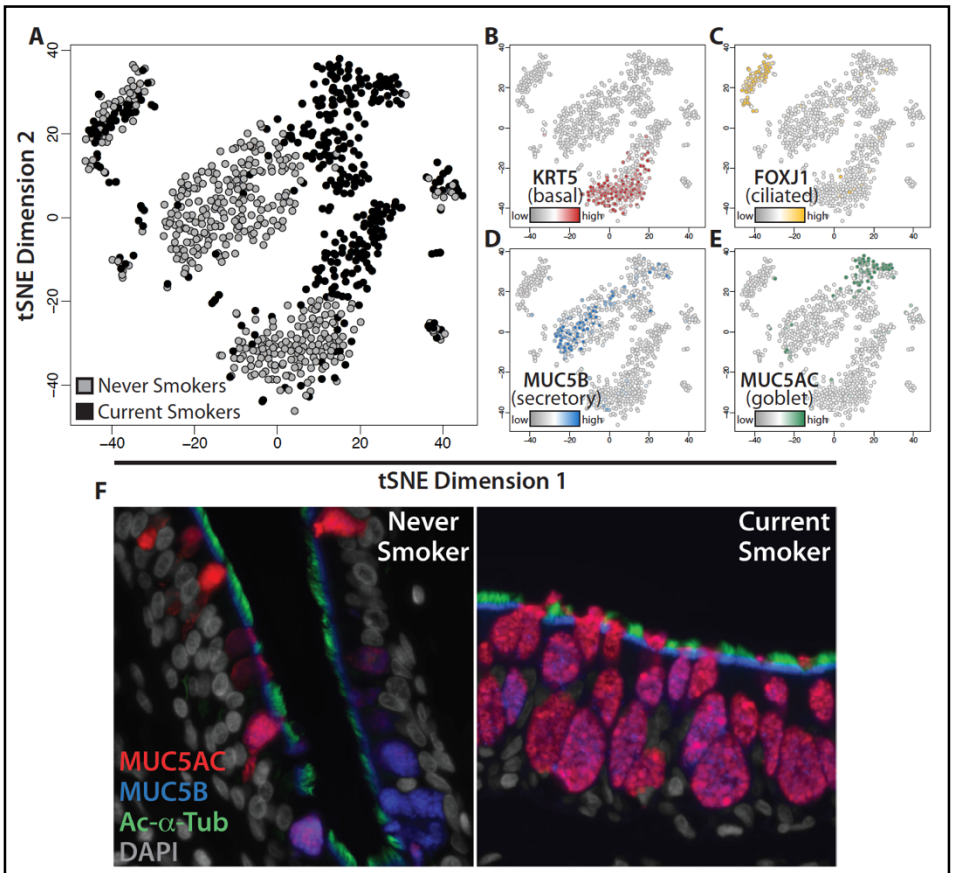

Figure 1. scRNA-seq of airway cells from never and current smokers. Dimensionality reduction was performed using t-SNE and relative expression of specific genes is displayed. Each point represents a single cell and cells with similar global gene expression patterns will be in similar locations on the plot. **(A).** Cells are highlighted according to smoking status (black = current smoker cells, grey = never smoker cells). **(B-E)** Cells are highlighted according to the expression levels of known airway epithelial marker cell type marker genes, specifically KRT5$^+$ basal, FOXJ1$^+$ ciliated, MUC5B$^+$ secretory, and MUC5AC$^+$ goblet. **(F)** Immunofluorescence microscopy was performed on bronchial airway tissue from an independent cohort of never and current smokers. An example of one never smoker (left) and one current (right) smoker are depicted showing fluorescent staining of MUC5AC (red), MUC5B (blue), and acetylated alpha tubulin (Ac-α-Tub, green, ciliated). Overall, these plots demonstrate that cells tend to cluster by both cell-type and by smoking status and that smoking can induce cell type shifts (increase in MUC5AC$^+$ cells, decrease in MUC5B$^+$ cells), a finding previously unappreciated by bulk profiling of this tissue.

scRNA-Seq Data of Airway Cells from Disease-free Never and Current Smokers Suggests the Presence of a Novel Cell Population and Cell Population-Specific Expression of Persistently Altered Genes in Former Smokers:

We have defined transcriptionally distinct cell populations using a machine learning algorithm called the "Latent Dirichlet Allocation" (LDA) that we are adapting into a single cell analysis method called celda. LDA can be used to decompose RNA-sequencing data into sets of co-expressed genes, referred to as "transcriptional states", which are then used to define corresponding sets of cells. This approach allows us to identify cellular subpopulations with distinct gene expression signatures. This analysis revealed 14 distinct "states", corresponding to 14 clusters of cells. Genes depicted in the **Figure 2** heatmap were previously identified as persistently up- (blue) or down-regulated (pink) in airways of former smokers, despite smoking cessation[7], along with known airway epithelial marker genes. Persistently down-regulated genes either specifically localized to never smoker MUC5B$^+$ secretory cells (e.g. MT1G) or were shared between never smoker KRT5$^+$ basal and MUC5B$^+$ cells (e.g. CX3CL1). A subset of persistently up-regulated genes localized to MUC5AC$^+$ goblet cells, as well as a novel smoking-associated population (e.g. CEACAM5). These novel smoking cells share features of basal (AQP3$^+$)[54] and goblet (SPDEF$^+$) cells, but lack expression of explicit marker genes, KRT5 (basal) and MUC5AC (goblet), indicating that they may be intermediate in nature. Additionally, some genes are up-regulated by all smoker cells (e.g. NQO1). Cell type specificity of genes persistently altered in former smokers suggests that smoking may permanently

alter the cellular constitution of the airways. In addition, some of the LC diagnostic genes up-regulated in LC airways (e.g. NKX3-1, **Figure 2**) are expressed by the novel smoking cells and goblet cells, suggesting that smoking-associated cell types changes are directly relevant to the lung cancer field of injury.

scRNA-Seq of Bronchial Brushings from Current and Former Smokers Undergoing Bronchoscopy for Suspected Lung Cancer Reveals Both Epithelial and Immune Cell Populations: Bronchial brushings were obtained from 8 former and 4 current smokers undergoing bronchoscopy for suspected LC (**Table 2**). After adjudication by Dr. Ehab Billatos, 7 subjects were determined to have lung cancer. For each donor, live cells were sorted into 96-well PCR plates and processed using the CEL-Seq2 RNA library preparation protocol. 192 cells/donor or 2,304 total cells were sequenced on an Illumina NextSeq500 in High-Output mode. At least 2,000 genes were detected and expressed on average for each cell. The relationship between cells is visualized using t-SNE (**Figure 3**). Expression of known cell type markers indicates that airway epithelial cell types identified in **Figure 1** were also profiled in this study (KRT5$^+$ basal, FOXJ1$^+$ ciliated, MUC5AC$^+$ goblet, MUC5B$^+$ secretory, ASCL3$^+$ neuroendocrine, as well as subpopulations of white blood cells: CD8$^+$ cytotoxic T cells, CD11C$^+$ monocytes, and CD22$^+$ mature B cells (naïve B cells markers, CD19/CD20 are not expressed well). These samples have an increased representation of the white blood cells because we modified the FACS sorting technique and did not sort on ALCAM and CD45 markers. The tissue obtained from bronchial brushings was treated with 0.25% Trypsin/EDTA for epithelial sheet dissociation and cells were sorted using a BD FACSAria II. FACS was used to isolate singlet events based on forward scatter height vs. forward scatter area (FSH-H vs. FSH-A). Dead cells (PI+) and red blood cells (GYPA/CD235a+) are stained and excluded, and live cells (Hoechst 33342+) are sorted. For each donor, single Hoechst 33342$^+$ PI$^-$ CD235a$^-$ cells were sorted into 96-well PCR plates, frozen on dry ice and stored at -80 °C until preparation for sequencing.

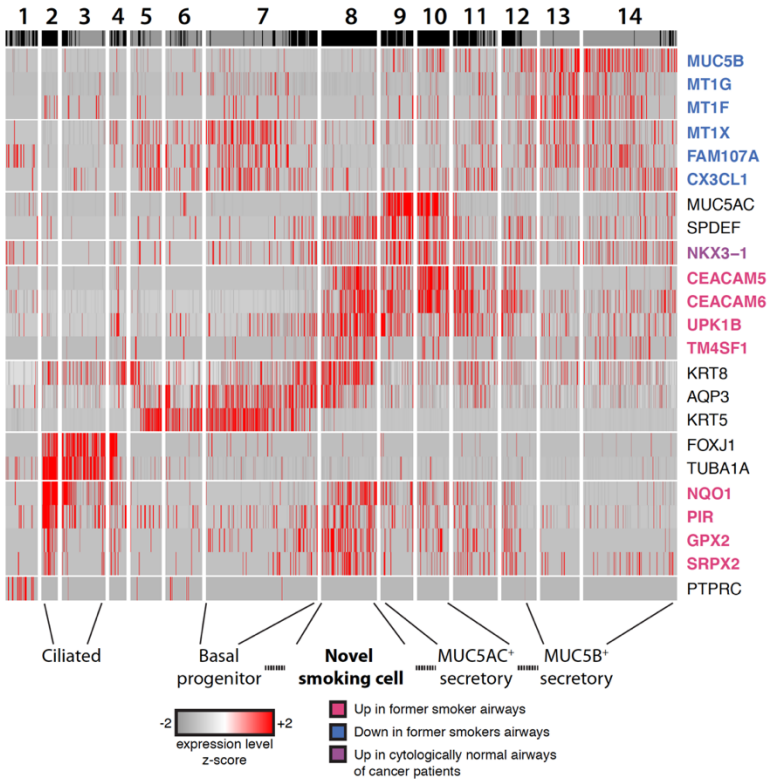

Figure 2. LDA Analysis illustrating cell transcriptional states. LDA analysis was performed across the scRNA-seq data from 1,140 airway cells collected from disease-free never and current smokers. Cells were partitioned into 14 clusters (separated by white columns) pertaining to 14 observed "transcriptional states" and select cell type marker genes as well as genes previously observed to be persistently altered in former smoker airways[7] are displayed by heatmap. Each column of the heatmap is a cell colored according to smoking status of the donor (grey = never, black = current smoker). Each row is a gene where z-score normalized gene expression is represented on a scale of grey (low expression) to red (high expression). Transcriptional state 8 characterizes the novel smoking cell population. Genes persistently up-regulated (pink) in former smokers localize to both the novel smoking cells and MUC5AC$^+$ goblet cells or are up-regulated across all cell types, whereas genes persistently down-regulated (blue) specifically localize to MUC5B$^+$ secretory cells or are shared by basal and MUC5B$^+$ cells.

Cancer Status	Smoking status	Sex	Age	Pack years
No Cancer (n=7)	4 C, 3 F	5 F, 2 M	66.7 (11.7)	38.6 (8.8)
Cancer (n=5)	5 F	3 F, 2 M	69.8 (9.5)	62 (53.9)

Table 2. Phenotypic information on smokers undergoing bronchoscopy for suspicion of lung cancer.

scRNA-Seq of Bronchial Brushings from Current and Former Smokers Undergoing Bronchoscopy for Suspected Lung Cancer Reveals Cell Type Specific Expression of LC-biomarker Genes:

scRNA-Seq of smokers with and without cancer reveals known populations of epithelial and immune cells. Some of these populations appear to have an enrichment of cells from either subjects with or without cancer (**Figure 4**). Subjects with cancer have an enrichment of basal cells whereas subjects without have an enrichment of mature B cells. Two genes, CH25H and VWA5A, previously described as lung cancer biomarker genes[13], localize to the basal cell and mature B cell populations. CH25H is involved in cholesterol and lipid metabolism and VWA5A is a putative tumor suppressor. The results suggest that biomarkers may be improved by isolating specific cell populations or by considering cell type-specific gene expression alterations.

Conclusions: We have successfully sequenced 3,456 cells obtained from bronchial brushings via bronchoscopy. Preliminary results reveal cell type shifts in response to smoke exposure along with a novel population of cells that express previously described genes that are irreversibly altered by smoking. Initial analyses of the scRNA-Seq from smokers with and without cancer shows both epithelial and immune cell populations and suggests that subjects with cancer may have more basal cells while no cancer subjects may have a depletion of mature B cells. In the next few months, we will be analyzing our cancer/no cancer scRNA-Seq dataset using the celda software developed by Dr. Campbell. These results will give us a high-resolution look at the different cell populations and we hope to understand the differences between former and current smokers, look for the existence of the novel cell type, and characterize whether or not it appears that there are tumor subtype specific differences. Finally, we plan to sequence additional cancer/no cancer subjects so that we can make robust conclusions.

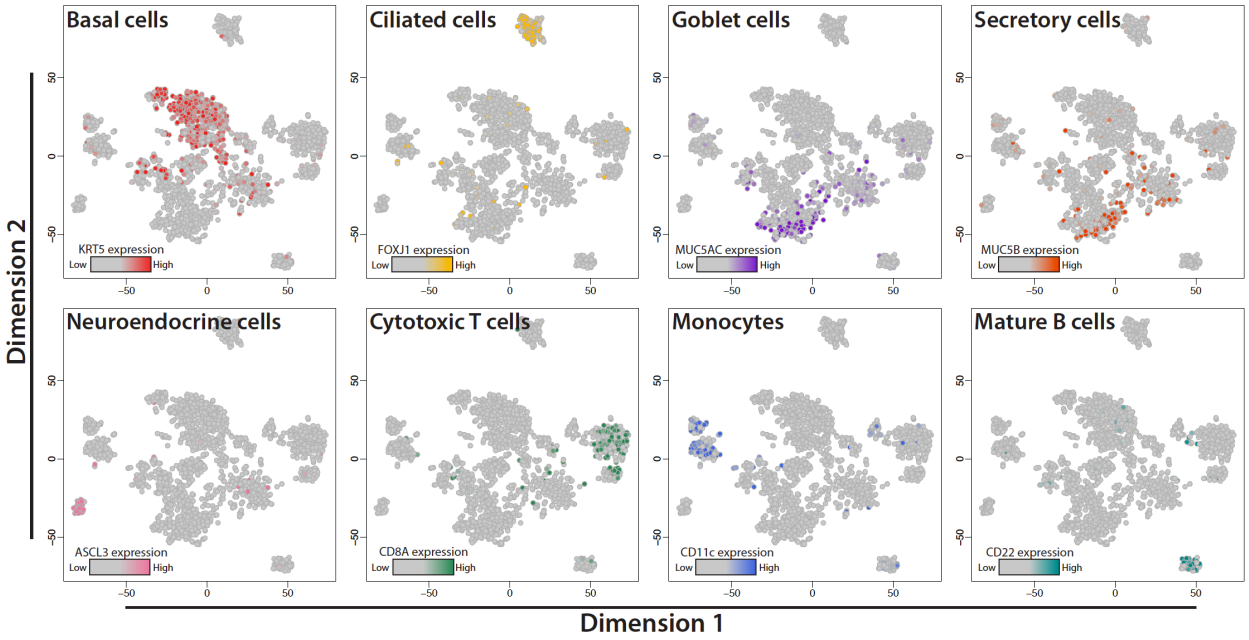

Figure 3. scRNA-seq of airway cells from smokers with and without lung cancer. We performed scRNA-Seq analysis on 2,304 cells from 12 subjects (7 subjects did not have lung cancer and 5 subjects were diagnosed with lung cancer). Dimensionality reduction was performed using t-SNE and relative expression of specific genes is displayed. Each point represents a single cell and cells with similar global gene expression patterns will be in similar locations on the plot. Cells are highlighted according to the expression levels of known airway epithelial marker cell type marker genes, specifically KRT5^{+} basal, FOXJ1^{+} ciliated, MUC5AC^{+} goblet, MUC5B^{+} secretory, ASCL3^{+} neuroendocrine, as well as subpopulations of white blood cells: CD8^{+} cytotoxic T cells, CD11C^{+}

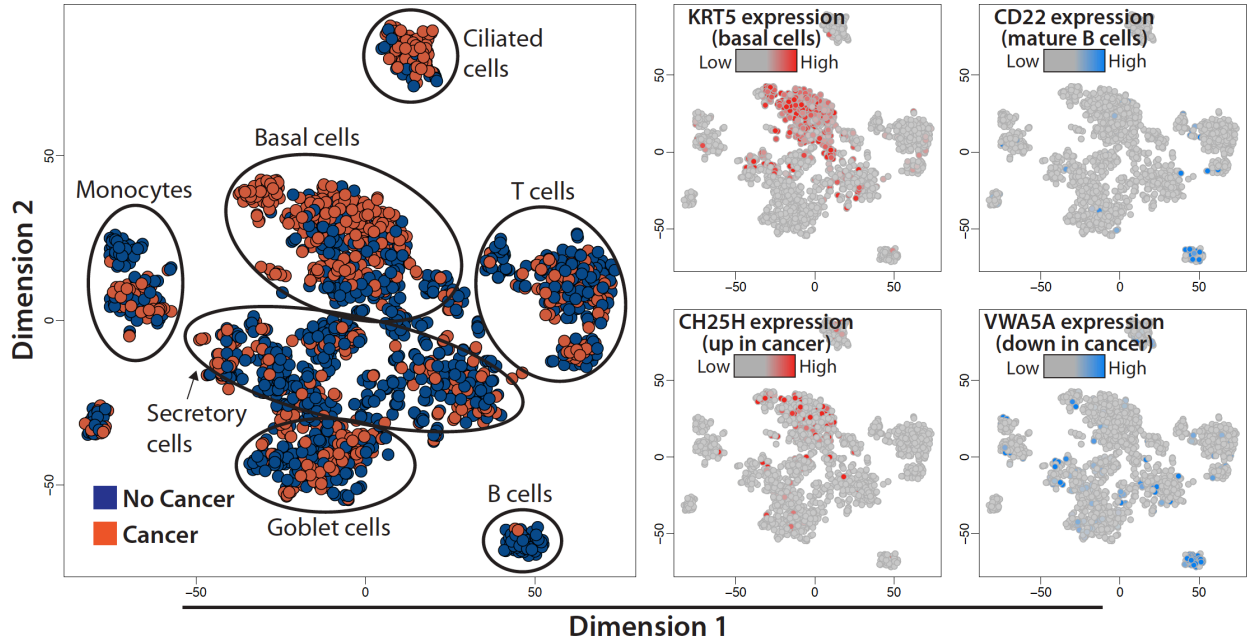

Figure 4. scRNA-seq of airway cells from smokers with and without lung cancer reveals cell type specific expression of lung cancer biomarker genes. Dimensionality reduction was performed using t-SNE and relative expression of specific genes is displayed. Each point represents a single cell and cells with similar global gene expression patterns will be in similar locations on the plot. Left panel. t-SNE plot with cells colored by cancer status. Middle and right panels. Select genes previously observed to be up- and down-regulated in the airways of lung cancer patients, CH25H and VWA5A, were found to localize to cells expressing the following cell type marker genes: KRT5^{+} basal cells and CD22^{+} mature B cells.

What opportunities for training and professional development has the project provided?
Grant Duclos, a graduate student, responsible for sorting the cells and preparing the libraries as part of the project has had the opportunity to study under Dr. Itai Yanai, Associate Professor at Technion – Israel Institute of Technology. Dr. Yanai is on sabbatical at the Broad Institute and has extensive experience in single cell sequencing and under his mentorship Grant has been able to effectively troubleshoot our protocols. Additionally, during the period of this grant, Joshua Campbell has become an Assistant Professor at BUMC in the Department of Medicine and Section of Computational Biomedicine. Using the knowledge he gained over the course of this project he started a Single Cell Working Group at BUMC that attracts researchers with diverse interests. The Single Cell Working Group is designed to educate the community about single cell sequencing and to discuss library preparations protocols and data analysis techniques. The goal is to develop a single cell sequencing analysis toolkit. In addition, these studies are leading to the development of a single cell tumor bank by the Cancer Center at BUMC.

How were the results disseminated to communities of interest?
The results of this work were presented at the Keystone Single Cell Meeting on May 27th, 2017 in Stockholm, Sweden, at the American Association of Cancer Research Meeting between April 1st and April 5th in Washington, D.C., and at the American Thoracic Society Annual Meeting between May 19th and 24th, 2017 in Washington, D.C. The references and abstracts for these presentations is below:

- Keystone Meeting: Don't fear the zeros: Identifying transcriptional states and cellular populations in sparse single-cell RNA-seq data with Bayesian hierarchical modeling. Sean Corbett, Zichun Liu, Tianwen Huan, Iris Yang, Grant Duclos, Jennifer Beane, W. Evan Johnson, Paola Sebastiani, Masanao Yajima, Joshua D. Campbell

  Background: Biology is divided into hierarchies: Complex tissues are composed of different cellular populations; each cell from each subpopulation contains a unique mixture of transcriptional states; and each transcriptional state is composed of groups of co-expressed genes. Single-cell RNA-seq can be used to explore these hierarchies by identifying all cellular populations within a sample and to determining the unique combination of transcriptional states that define each subpopulation. However, single cell RNA-seq data is noisy and contains many zeros due to the challenges inherent in amplifying small amounts of RNA.
  Methods: With these goals and challenges in mind, we implemented Bayesian hierarchical models that reflect the hierarchies observed in biological systems. These models can be used to cluster co-expressed genes into transcriptional states, cells into subpopulations, and quantify the proportion of each cellular subpopulation within independent samples. Importantly, these models can handle sparse count-based data without additional normalization.
  Results: We applied these models to single cell RNA-seq data generated from airway epithelial cells from smokers and non-smokers. Cell-type specific gene-expression alterations were induced in the airway epithelium of smokers. Additionally, a novel subpopulation was discovered in the airway of smokers that did not expression known cell-type markers and likely represents cells transitioning from a progenitor state to a secretory state.
  Conclusions: Overall, these models represent novel approaches to characterizing cellular and transcriptional heterogeneity in biological samples using single-cell RNA-seq data.

- ATS: Single Cell RNA Sequencing Reveals Smoking-Associated Alterations in Bronchial Airway Epithelial Subpopulations. Grant E. Duclos, Joshua D. Campbell, Yaron Gesthalter, Patrick Autissier, Yves M Dumas, Robert Terrano, Gang Liu, Marc E Lenburg, Avrum Spira, Jennifer Beane

  RATIONALE: We have previously shown that bronchial airway epithelial gene expression reflects the physiologic response to cigarette smoke exposure. In this study, we use single cell mRNA sequencing to profile the transcriptomes of individual bronchial epithelial cells from current and never smokers in order to detect smoking-associated alterations within specific epithelial cell types and to discover novel subpopulations in the airways of smokers.
  METHODS: We obtained bronchial brushings from current smokers (n=6) and never smokers (n=6) and isolated single cells by FACS. The CEL-Seq RNA library preparation protocol was used

to sequence the transcriptomes of 1,140 cells (n=95/donor). Latent Dirichlet allocation was used to identify cellular subpopulations and transcriptional states.
RESULTS: Distinct populations of bronchial cells expressed known markers of basal (*KRT5*), ciliated (*FOXJ1*), secretory (*SCGB1A1* and *MUC5AC*), and neuroendocrine (*ASCL3*) epithelial cells, as well as natural killer white blood cells (*NKG7*). In the airways of smokers, we observed an increase in the abundance of *MUC5AC*$^+$ secretory cells as well as a decrease in the abundance of *KRT5*$^+$ basal cells and *SCGB1A1*$^+$ secretory cells. A novel subset of KRT8$^+$ cells that lacked expression of other known cell type markers was identified in the airways of smokers and may represent a population previously described as undifferentiated intermediate cells. Genes involved with metabolism of polycyclic aromatic hydrocarbons (e.g., *CYP1B1*) were detected in smoker secretory cells, whereas genes involved in the metabolic response to cigarette smoke toxins such as aldehydes (e.g., *ALDH3A1*) and quinones (e.g., *NQO1*), were most highly expressed by smoker ciliated cells. Interestingly, the novel KRT8$^+$ cells identified in smokers expressed genes known to promote *MUC5AC*$^+$ secretory cell differentiation (e.g., *SPDEF*), but did not express *MUC5AC* itself, suggesting that these may be pro-*MUC5AC*$^+$ secretory intermediate cells.
CONCLUSION: Cell type-specific transcriptomic alterations and shifts in epithelial cell population abundance were observed in smoker airways. The findings from this study provide insights into the molecular response to smoking in subpopulations of cells within the bronchial epithelium and may help identify how different cell populations contribute to the pathogenesis of tobacco-related lung diseases.

- AACR: Single Cell RNA Sequencing Reveals Smoking-Associated Alterations in Bronchial Airway Epithelial Subpopulations. Grant E. Duclos, Joshua D. Campbell, Yaron Gesthalter, Patrick Autissier, Yves M Dumas, Robert Terrano, Gang Liu, Marc E Lenburg, Avrum Spira, Jennifer Beane

  RATIONALE: We have previously shown that bronchial airway epithelial gene expression reflects the physiologic response to cigarette smoke exposure. We have also shown that gene expression differences in cytologically normal airways cells can serve as a diagnostic biomarker for lung cancer. In this study, we use single cell RNA-seq to profile transcriptomes of individual bronchial epithelial cells from current and never smokers in order to detect smoking-associated alterations within specific epithelial cell types and to discover novel subpopulations that develop as a result of smoke exposure. This approach may be useful for identifying cell type-specific transcriptomic changes in the airways of cancer patients, which may lead to a better understanding of lung carcinogenesis and new approaches to early lung cancer detection.
  METHODS: We obtained bronchial brushings from current smokers (n=6) and never smokers (n=6) and isolated single cells by FACS. The CEL-Seq RNA library preparation protocol was used to sequence the transcriptomes of 1,140 cells (n=95/donor).
  RESULTS: Distinct populations of bronchial cells expressed known markers of basal (*KRT5*), ciliated (*FOXJ1*), secretory (*SCGB1A1*, *MUC5AC*) epithelial cells, as well as white blood cells (*CD45*). In the airways of smokers, we observed an increase in abundance of *MUC5AC*$^+$ secretory cells as well as a decrease in abundance of *KRT5*$^+$ basal cells and *SCGB1A1*$^+$ secretory cells. A novel subset of KRT8$^+$ cells that lacked expression of other known cell type markers was identified in the airways of smokers and may represent a population previously described as undifferentiated intermediate cells. Genes involved with metabolism of polycyclic aromatic hydrocarbons (*CYP1B1*) were detected in smoker secretory cells, whereas genes involved in the metabolic response to cigarette smoke toxins such as aldehydes (*ALDH3A1*) and quinones (*NQO1*), were most highly expressed by smoker ciliated cells. Interestingly, the novel KRT8$^+$ cells identified in smokers expressed genes known to promote *MUC5AC*$^+$ secretory cell differentiation (*SPDEF*), but did not express *MUC5AC* itself, suggesting that these may be pro-*MUC5AC*$^+$ secretory intermediate cells. Furthermore, we found that genes previously associated with higher expression in the airways of lung cancer patients were enriched among genes most strongly associated with smoker ciliated and secretory cells, whereas genes with lower expression in lung cancer were enriched among genes most strongly associated with white blood cells.

CONCLUSION: We have identified cell type-specific transcriptomic alterations and shifts in epithelial cell population abundance in smoker airways. In future studies, profiling the transcriptomes of single cells from bronchial airways of smokers with and without lung cancer may lead to the identification of specific cellular subpopulations contributing to the airway field of lung-cancer associated injury.

What do you plan to do during the next reporting period to accomplish the goals?
We plan to pursue an in-depth analysis of the new scRNA-Seq data that we have generated across the subjects with and without cancer. Additionally, we are writing up a manuscript on the scRNA-Seq data we generated across the healthy never and current smoking subjects. We also plan to sequence both healthy formers smokers and additional smokers with and without cancer.

IMPACT:
What was the impact on the development of the principal discipline(s) of the project?
To date, human bronchial epithelial cells obtained via bronchoscopy have not been sorted and processed for single cell RNA sequencing. The cells need to be immediately taken from the bronchoscopy suite and processed for FACS sorting. FACS sorting of the cells is accomplished as stated above and the sorted cells are frozen in 96-well plates. The CEL-Seq RNA preparation protocol has been modified to provide the ability to process hundreds of cells from several subjects. All of the methodology and protocols developed as part of this project will directly benefit other groups that are attempting to examine single cell transcriptomics in human clinical samples. In addition, all microarray or RNA sequencing studies performed to date on human airway epithelial cells have been conducted using a bulk population of cells. Expression values represent the mean behavior of all the cells profiled in a given sample and do not capture cell-to-cell gene expression variation. Using a single cell approach, we are able to characterize the cell types (both known and novel) that are present and begin to identify the cell types responsible for the lung cancer-specific airway field of injury.

Using the scRNA-Seq data generated as part of this project we have applied for early lung cancer detection Stand Up To Cancer award (PI, Avrum Spira) and an NIH R01 (PI, Jennifer Beane) grant. We have also obtained industry sponsored funding to continue single cell sequencing projects to examine bronchial premalignant disease. Additionally, we have started to use the cell type specific transcriptional profiles to deconvolute airway epithelium datasets that we have previously profiled using bulk tissue.

What was the impact on other disciplines?
The project is multidisciplinary and will impact the study of lung cancer and epithelial cell biology as well as contribute to molecular biology and bioinformatics methods. The current data presented above suggests new markers to study epithelial cell types/subpopulations that appear to be important in the process of lung carcinogenesis. Additionally, the application of computational techniques such as LDA that are being developed as part of this project will be useful to researchers analyzing other single cell datasets.

What was the impact on technology transfer?
Nothing to report.

What was the impact on society beyond science and technology?
A gene expression based biomarker for improving lung cancer diagnosis known as PERCEPTA^{TM} and commercialized by Veracyte (http://www.veracyte.com) is currently available following promising clinical trial results[10]. The test helps identify patients at low risk for having lung cancer after a non-diagnostic bronchoscopy ordered as a result of CT scan abnormalities. The findings in this project may help develop an improved diagnostic test that will impact the clinical management of high lung risk patients.

CHANGES/PROBLEMS:
Changes in approach and reasons for change
When we generated our scRNA-Seq across the healthy current and never smoking subjects, for each 96-well plate, we sorted 85 ALCAM^{+} epithelial cells and 10 CD45^{+} immune cells (1 well is left empty as a

negative control). For the generation of the cancer/no cancer set, we took an unbiased approached and sorted all live cells into 96-well plates. As a result, the data (**Figure 4**) reflects a higher representation of immune cells. As the immune system may play an important role in maintaining a healthy airway epithelium, we hope that his new approach, will allow us to understand how the immune cell populations change with smoke exposure and the presence of cancer. Additionally, the single cell RNA library preparation procedure was modified to incorporate recently published changes to the "CEL-Seq" protocol[14], now referred to as "CEL-Seq2". In brief, changes primarily involve incorporation of the 3' sequencing adaptor via random hexamer-based reverse transcription (rather than RNA ligation). Implemented changes improved protocol efficiency and sensitivity.

Actual or anticipated problems or delays and actions or plans to resolve them
Our ability to collect samples was impacted by an IRB audit that revealed problems with our sample collection. During the audit and while we rectified all issues with the IRB and the DoD HRPO we suspended recruitment of subjects. All problems have been resolved and we now have approval to continue collecting samples after about a 9-month delay. Collection of samples at Boston University Medical Center is slow and there are other studies competing for the same subjects. Nevertheless, we hope over the next year to collect at least 12 more cancer/no cancer subjects.

Changes that had a significant impact on expenditures
Nothing to Report.

Significant changes in use or care of human subjects, vertebrate animals, biohazards, and/or select agents
Nothing to report.

Significant changes in use or care of human subjects
Nothing to report.

Significant changes in use or care of vertebrate animals.
Nothing to report.

Significant changes in use of biohazards and/or select agents
Nothing to report.

PRODUCTS:
Publications, conference papers, and presentations
See abstracts above.

Journal publications.
Nothing to report.

Books or other non-periodical, one-time publications.
Nothing to report.

Other publications, conference papers, and presentations.
Nothing to report.

Website(s) or other Internet site(s)
Nothing to report.

Technologies or techniques
The techniques, including cell sorting, library preparation, and analysis methods developed in this project will be shared through publication of a manuscript reporting the findings of single cell sequencing experiments profiling smokers.

Inventions, patent applications, and/or licenses

Nothing to report.

Other Products
Nothing to report.

PARTICIPANTS & OTHER COLLABORATING ORGANIZATIONS
What individuals have worked on the project?

Name:	*Jennifer Beane-Ebel*
Project Role:	*Principal Investigator*
Researcher Identifier (e.g. ORCID ID):	
Nearest person month worked:	*3.6*
Contribution to Project:	Overseeing all aspects of the project including sample collection, experimental design, and data analysis
Funding Support:	NIH/NCI, Industry, Internal Funds

Name:	*Joshua Campbell*
Project Role:	Co-investigator
Researcher Identifier (e.g. ORCID ID):	
Nearest person month worked:	1.2
Contribution to Project:	Overseeing all aspects of the project with the PI and leading the computational analysis of the data
Funding Support:	LUNGevity Foundation, Industry, Internal Funds

Name:	*Denise Fine*
Project Role:	Research Coordinator
Researcher Identifier (e.g. ORCID ID):	
Nearest person month worked:	0.24
Contribution to Project:	Consenting patients and sample collection and supervision of IRB team.
Funding Support:	Donated Time

Name:	*Grant Duclos*

Project Role:	Graduate Student
Researcher Identifier (e.g. ORCID ID):	
Nearest person month worked:	12.0
Contribution to Project:	Collecting samples, sorting cells, and preparing RNA sequencing libraries from the cells, and data analysis
Funding Support:	NIH/NHLBI Training Grant

Name:	*Ehab Billatos*
Project Role:	Research Coordinator
Researcher Identifier (e.g. ORCID ID):	
Nearest person month worked:	0.24
Contribution to Project:	Adjudication of cancer status via chart review
Funding Support:	Pulmonary Fellow

Name:	*Jian You*
Project Role:	Technician
Researcher Identifier (e.g. ORCID ID):	
Nearest person month worked:	8.52
Contribution to Project:	Library Preparation
Funding Support:	DoD

Has there been a change in the active other support of the PD/PI(s) or senior/key personnel since the last reporting period?

J. Beane, PI
New Research Support
Sponsored Research Agreement with Industry (A. Spira, PI)

J. Campbell, Co-I
New Research Support
LUNGevity Career Development Award

What other organizations were involved as partners?
Organization Name: Broad Institute

Location of Organization: Boston, MA
Partner's contribution to the project: Collaboration. We have been collaborating with Itai Yanai, Associate Professor at Technion – Israel Institute of Technology during his sabbatical at the Broad Institute to help develop our single cell sequencing library preparation protocol.

SPECIAL REPORTING REQUIREMENTS
None.

APPENDICES:

References

1. Franklin, W. A. *et al.* Widely dispersed p53 mutation in respiratory epithelium. A novel mechanism for field carcinogenesis. *J. Clin. Invest.* **100,** 2133–2137 (1997).
2. Wistuba, I. I. *et al.* Molecular damage in the bronchial epithelium of current and former smokers. *J. Natl. Cancer Inst.* **89,** 1366–1373 (1997).
3. Tang, X. *et al.* EGFR tyrosine kinase domain mutations are detected in histologically normal respiratory epithelium in lung cancer patients. *Cancer Res.* **65,** 7568–7572 (2005).
4. Mao, L. *et al.* Clonal genetic alterations in the lungs of current and former smokers. *J. Natl. Cancer Inst.* **89,** 857–862 (1997).
5. Powell, C. A., Klares, S., O'Connor, G. & Brody, J. S. Loss of heterozygosity in epithelial cells obtained by bronchial brushing: clinical utility in lung cancer. *Clin. Cancer Res.* **5,** 2025–2034 (1999).
6. Spira, A. *et al.* Effects of cigarette smoke on the human airway epithelial cell transcriptome. *Proc. Natl. Acad. Sci. U.S.A.* **101,** 10143–10148 (2004).
7. Beane, J. *et al.* Reversible and permanent effects of tobacco smoke exposure on airway epithelial gene expression. *Genome Biol.* **8,** R201 (2007).
8. Spira, A. *et al.* Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nat. Med.* **13,** 361–366 (2007).
9. Beane, J. *et al.* A prediction model for lung cancer diagnosis that integrates genomic and clinical features. *Cancer Prev Res (Phila)* **1,** 56–64 (2008).
10. Silvestri, G. A. *et al.* A Bronchial Genomic Classifier for the Diagnostic Evaluation of Lung Cancer. *N. Engl. J. Med.* (2015). doi:10.1056/NEJMoa1504601
11. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep* **2,** 666–673 (2012).
12. Bonser, L. R., Zlock, L., Finkbeiner, W. & Erle, D. J. Epithelial tethering of MUC5AC-rich mucus impairs mucociliary transport in asthma. *J. Clin. Invest.* **126,** 2367–2371 (2016).
13. Whitney, D. H. *et al.* Derivation of a bronchial genomic classifier for lung cancer in a prospective study of patients undergoing diagnostic bronchoscopy. *BMC Med Genomics* **8,** 18 (2015).
14. Hashimshony, T. *et al.* CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17,** 77 (2016).